图书情报工作
第 64 卷 第 9 期 2020 年 5 月

学术论文研究亮点的语言学特征与分布规律研究

■ 索传军 于果鑫
中国人民大学信息资源管理学院 北京 100872

摘 要：[目的/意义] 论文出版过程中，若能够合理有效地呈现出一篇学术论文的核心观点，不仅可以大大减少科研人员在查找和筛选文献上花费的时间，而且有助于阅读与理解。[方法/过程] 通过标注 385 篇 XML 格式期刊论文，构建了研究语料库，再利用关键词分析法对亮点的语言学特征进行分析，借助自然语言处理算法探索亮点的分布特征。[结果/结论] 亮点是一组规范的、语义明确的短句的集合，是一篇学术论文与其他论文相比较的新观点、新视角、新方法、新思路、新结果、新结论等内容的体现。亮点具有新颖性、简明性、易读性、“宣传”性等特点。亮点分为研究创新型亮点、研究方法型亮点、研究过程型亮点与研究结论型亮点，本文发现了亮点在正文和各章节的分布情况。

关键词：学术论文 研究亮点 亮点价值 语言学特征 分布特征
分类号：G210.7
DOI：10.13266/j.issn.0252 – 3116.2020.09.012

1 引言

如何科学有效地呈现一篇学术论文的核心观点，促进人们对学术论文内容的快速了解，节省人们发现和阅读文献的时间，已成为亟待解决的问题。爱思唯尔于 2010 年正式提出学术论文研究亮点（Research Highlights）。其认为，亮点是一组论文的核心发现，是由论文作者在投稿时自主编写提交，用以帮助用户快速了解论文的 3 – 5 个要点[1]。但关于亮点的内涵和外延，以及意义和特征等并没有说明，学术界和出版界也没有形成统一认识，一定程度上影响了亮点概念和应用的普及，影响了以亮点为基础的论文语篇语义的探索。

通过研究发现，关于亮点的研究较少。香港学者 P. Tse[2] 指出，作为研究论文的伴随内容，如亮点，可以支撑清晰的学术立场和可信的学术形象。台湾作者 W. Yang[3] 通过对 240 篇期刊论文亮点的语言学分析，探究了亮点的评价性语言与交互式语篇的特点，并采用问卷法调查了编辑和作者对亮点的看法。相对于亮点而言，国内外学者对学术论文的创新点等内容的研究较多，F. Ronzano[4]、T. Dahl[5]、B. Fisas[6]、温有奎[7]、毛琛瑜[8] 等学者借助自然语言处理和机器学习等技术分别从不同视角对学术论文中的新发现和重要结论进行了识别和抽取等实践探索。然而，研究亮点，既不同于创新点，也不同于要点，它们之间既存在一定联系，又有区别。对其进行研究，不仅有利于改善编辑对学术论文创新性的判断和学术价值的评价，更有利于读者发现和获取自己所需的论文及其重要内容。因而，对学术论文亮点的研究具有重要的意义和价值。

2 亮点自动抽取的意义与价值

2.1 亮点自动抽取的意义

通常，一篇学术论文的意义和价值，只有阅读之后才能够判断。然而，由于论文数量较多，读者没有足够的时间去阅读和选择，往往会错失一些较有价值的文献。多年来，知识工程领域利用计算机自然语言处理技术一直研究“自动摘要”的编写，但没有取得良好的效果。为此，爱思威尔提出，让作者标注研究亮点，希望缓解这一矛盾。事实上，人工标注研究亮点，准确度较高，但成本高，效率低，无法解决海量存量论文亮点标注的需要。因而，通过对亮点的语言学特征，及其在论文中分布规律的探寻，实现对论文的亮点自动识别和抽取。标明每篇论文的研究亮点，既可以节省读者的时间，又能促进其快速传播，具有较大的意义和价值。

作者简介：索传军（ORCID：0000 – 0002 – 7416 – 1513）：教授，博士生导师；于果鑫（ORCID：0000 – 0002 – 7787 – 8511）：硕士研究生，通讯作者，E-mail：18660113368@163.com。
收稿日期：2019 – 06 – 13 修回日期：2019 – 10 – 28 本文起止页码：104 – 113 本文责任编辑：杜杏叶

2.2 亮点的定义

亮点从语义内容看,属于学术论文,从表现形式看,属于文本。学术论文包含有摘要、正文、参考文献等,文本包含有长文本与短文本。对比可知,内容上,亮点与摘要有相似之处;形式上,亮点与短文本相似。亮点与摘要相比,更加新颖、简洁;亮点与短文本相比,具有更加简短、规范和语义明确等特点。

通过上述分析可知:亮点是一组规范的、语义明确的短句集合。包含着五方面的含义:①亮点必须符合语法规范,且语义完整;②亮点在表达充分的基础上,应当尽可能简短;③亮点是关于学术论文某一方面新颖性的说明;④亮点表达论文的重要内容,体现一篇论文的独特之处;⑤亮点可以让读者对论文创新性内容有一个概览的了解。

2.3 亮点的特点

本文通过对论文亮点的分析发现,亮点具有新颖性、简明性、易读性、"宣传"性(Promotional[3])等特点。

(1)亮点内容的新颖性。亮点是一篇论文新观点、新视角、新方法、新思路、新结果、新结论等重要内容的体现,因此新颖性是亮点最为基本的特点。具体而言,亮点的内容必须由作者独自创作完成,比既有研究成果更加新颖,可以是改进和修正,也可以是颠覆与突破。

(2)亮点表达的简明性。亮点前置于摘要呈现给读者,其简明性无疑是关键因素。爱思唯尔规定每条亮点不超过 85 个字符,这就要求作者需要在不影响读者理解的前提下,使用尽可能少的字符将论文的重要内容充分表达。倘若亮点语言不够精炼,字数过多,则导致亮点在形式上无异于摘要,并且可能会给读者阅读造成障碍。

(3)亮点对读者的易读性和"宣传"性。亮点可以吸引潜在读者阅读全文,因此亮点对读者具有易读性和"宣传"性。易读性主要体现在亮点表述的通俗易懂,不同于对公式、数据的冷冰冰地罗列,亮点是学术观点的生动表达。易读性使读者在阅读亮点时不需要对该课题有过多的知识背景,就可以快速掌握该论文的核心内容。另一方面,大量使用"加强语"(Intensifier)是学术语篇中宣传论文的一个特征[9],作者在编写亮点时通常使用副词或形容词等"加强语"来强化自己的观点,这决定了亮点的"宣传"性。

2.4 亮点的价值

一篇学术论文的亮点既有利于编辑对稿件价值的判断,也有利于读者对论文的选择,还有利于作者对论文核心观点的宣传。亮点对于读者、审稿编辑、期刊出版商和作者均有重要意义。具体表现在以下几个方面:

(1)有利于更高效地进行论文审稿,助力学术监审。亮点可以帮助编辑与审稿专家对论文的学术价值进行初步的判断,加快审稿速度,提高评审效率。囿于作者核心研究观点呈现不合理、不清晰,专家对论文核心观点的把握与提取成为审稿的焦点和难点[10]。因此,有学者提出若由作者自身明确标注出论文的核心内容,便可以加快审稿编辑对来稿的学术不端性、创新性和价值进行初步判断。

(2)有利于提高学术出版商论文的吸引力,拓展增值效益。"出版或毁灭"(Publish or Perish)的学说发布以来,学术出版界逐渐成为了一个高度竞争的领域,学术期刊出版商们争相努力吸引潜在的作者并扩大它们的读者群。亮点的简明与便捷使之成为提升学术出版商竞争力的有力抓手,此外,亮点还可以吸引潜在的读者购买完整的访问权限,创造增值效益。

(3)有利于读者对论文价值的判断,提高阅读效率。论文发表的重要目的是让读者学习、借鉴和利用创新性成果并进行知识的再创造[11]。不过,要了解一篇论文的核心观点需要阅读大量文字、耗费较长时间。亮点无疑会大大节省读者的时间和精力,并提升论文的"感知程度",因此亮点可以帮助读者加快论文的选择。

(4)有利于论文作者宣传自己的论文,传播学术观点。作者直接呈现研究的主要发现、观点和成果,有助于读者选择,加快论文的传播和利用,提升作者在学界的影响力。

3 研究语料库建设

3.1 数据来源

本文通过爱思唯尔 ScienceDirect 电子期刊数据库获取 *International Journal of Information Management* 期刊 2016 年至 2018 年共 385 篇全文数据,并运用 Oxygen XML Editor 软件,依照一定的标注规则对论文进行标注,构成本研究的数据来源。

3.2 XML 文本标记规则

根据一篇完整的期刊论文的结构特点和 XML 可扩展标记语言的语法特点,结合本研究需要,创建自定义标记规则(见表 1)。其中包含题录信息、亮点、摘要、关键词以及正文等内容,可以实现一篇学术论文的完整标记。

图书情报工作
第 64 卷 第 9 期　2020 年 5 月

表 1　文本标记规则及示例

序号	类别	标记语言	示例
1	全文	<publication>	<publication> …… </publication>
2	期刊名称	<journal>	<journal> International Journal of Information Management </journal>
3	出版时间	<time>	<time> Volume 36, Issue 6, Part A, December 2016, Pages 1062 – 1074 </time>
4	论文题目	<title>	<title> Information technology resource, knowledge management capability, and competitive advantage: The moderating role of resource commitment </title>
5	作者	<author>	<author> Hongyi Mao </author> <author> Shan Liu </author>
6	DOI 号	<doi>	<doi> https://doi.org/10.1016/j.ijinfomgt.2016.07.001 </doi>
7	亮点	<highlight>	<highlight id = "1"> IT resources positively affect knowledge management capability (KMC). </highlight> <highlight id = "2"> Resource commitment positively influences KMC. </highlight>
8	摘要	<abstract>	<abstract> <![CDATA[…… ……]]> </abstract>
9	关键词	<keywords>	<keywords> Information technology resource </keywords> <keywords> Knowledge management capability </keywords> <keywords> Resource commitment </keywords> <keywords> Resource-based view </keywords>
10	正文	<section>	<section name = "Introduction category = "introduction"> <![CDATA[…… ……]]> </section>

说明:①作者、关键词和亮点需要逐条标记,亮点需要编号加以区分。②"<![CDATA[ ]]>"用于表示特殊符号。③从引言开始的正文标记,需要加上"name"和"category"两个属性,"name"为作者表述的标题名称;"category"为规范名称。例如,语句 <section name = "literature review" category = "background">,说明论文作者使用"literature review"表达文中的这部分内容,而在本研究归纳的统一标准化论文结构中,应该采用"background"来标注

3.3　亮点标记规则

亮点的标记是本研究的一个关键问题。本文根据句子匹配、短语相关、内容相关等原则找出每个亮点在全文出现的位置,并做出相应标记。由于许多亮点在论文中并不以原始语言出现,换言之,亮点是作者根据论文中具体内容归纳而成,因此简单的字句匹配并不准确,需要根据内容逐句甄别。本研究采用如下方法对亮点进行标记。

(1)首先对"highlight"标签元素添加编号属性"highlight id = 1,2,3……",例如:语句 <highlight id = "1"> IT resources positively affect knowledge management capability (KMC). </highlight> 表示第一条亮点。"target"标签用于对应某条亮点,"match"标签表示全文中的"highlight"语句与亮点的匹配情况。标记时要注意全文标记处前后分别增加"]]>"和"<![CDATA["两个符号(见图 1),其意义在于提前结束段落前的"<![CDATA[",否则标记符号会被转译成普通字符。

(2)其次,若一个句子仅与某条亮点的一部分内容相匹配,或者多个句子描述一条亮点,则"match"标签标记为"part",意为部分匹配。若描述了"highlight"的全部内容,则标记为"full",意为完全匹配。

```
<abstract><![CDATA[The design and utilization of Electronic Process
Guides (EPGs) have been studied in Software Engineering (SwE) since
the 1990s. However, the empirical findings from surveys, case
studies, and experiments on the beneficial effects of their
utilization are still lacking. Thus, we suggest that further research
on the utilization of EPGs is required. ]]><h target="1"
match="part">In this study, we are interested in gaining insights on
the effects of using EPGs on objective metrics (learning score, time
effort) and subjective metrics (perceived usefulness, ease of use,
and value), by comparing three EPG designs (a simple PDF-based EPG, a
normal HTML-based EPG, and a sophisticated Java-based EPG) with
different blocks of experimental subjects (practitioners,
academicians, novices, and experts).</h><![CDATA[ To this end, we
have conducted a controlled experiment with a sample of international
participants in the domain of IT Service Management. We found that
```

图 1　亮点标记示例

(3)再次,亮点是作者的概括性内容,若段落中大部分内容描述一条亮点,则整个段落都应标记。若正文某个段落描述了一条亮点,该段落中某个句子 S1 描述另外一条亮点,为了避免嵌套,该段落中 S1 单独标注,前后两部分"match"分别标为"part"。

（4）最后，一条亮点可能出现在文中多个地方，或者文中多个地方都有内容与一则亮点匹配的语句、段落，那么需要将其全部标出（见图2）。此外，下列特殊情况不做标记：作为问题出现的亮点不予标记；作为假设出现的亮点不予标记；若亮点与引用句子观点一致，引用句子不予标记。

Highlights

- Task-related features affect perceived usefulness.
- Mood-related features affect perceived enjoyment.

Fig. 2. Research model.

3.1.]]> <h target="1;2" match="full;full">Effects of TR and MR features on perceived usefulness and perceived enjoyment</h> <![CDATA[

图 2 “一对多”亮点标记方法

4 亮点的语言学特征分析

学术论文研究亮点的语言学特征主要体现在特征词（关键词）和常用表达方式两方面。因而，本文的研究思路是，首先，进行关键词分析，并根据关键词词义对亮点分类，然后再汇总各类型亮点的常用表达方式。

4.1 关键词分析

关键词分析法作为定性分析的一种形式，可以帮助识别给定学术语篇中词汇的重要性，并且有助于建立对语篇中一般意义词语的搭配和合成关系的清晰理解。关键词是指与参考语料库其他词相比，出现频率较高的词[12]。高频词通常可以分成“高频通用词”和“高频特征词”两类。高频通用词是语言中通用的高频词。高频特征词是反应特定内容特点和风格的高频词。本文的关键词即为文本中的“高频特征词”。

除此之外，亮点中高频特征词的甄选和分析至关重要。表示过程的动词和表示物体的名词是人类在认知世界中划分出来的两大范畴[13]。动词在许多语法理论里被看成句子结构最重要的成分，动词决定句子的基本结构，也是句子的核心[14]，而其中言语行为动词（Speech Act Verbs）构成动词词汇的一个重要部分，言语行为动词对领悟人与人、人与事物之间的关系非常重要。

从认知角度看，言语行为动词所构成的句子都会涉及两个参与者：陈事者与客体。客体可以是具体的事物，如“I told a thing”句中的“a thing”，也可以是抽象的事物，例如亮点中提出的某一观点。经调查发现，在亮点句中的陈事者主要是论文作者或论文本身，例如“The study identifies the risk of BDT.”，其中，“the study”是陈事者，“the risk of BDT”是客体，而言语行为动词“identify”表示了两点的核心语义——“识别出”。这说明了言语行为动词作为关键词的合理性。概括地说，本文的关键词是指文本中的言语行为动词。

4.2 关键词词频统计

本研究使用 WordSmith Tools 的关键词检索程序，来确定亮点文本的关键词以及它们在文本中的位置。首先使用 WordList 工具建立两个单词列表，一个是根据要考察的文本建立的亮点文本，而另一个作为参照的单词列表，是根据较大型的由同类文本组成的语料库建立的，参照单词列表将为比较提供背景数据。

基于关键词检索程序的文本内容分析，总共选出154 个言语行为动词用于进一步研究，再借助 WordSmith Tools 中的协调函数显示亮点的完整条目，从而确认亮点为有效亮点。通过对关键词词频的归纳和统计，得出部分高频言语行为动词关键词词频统计表，如表 2 所示：

表 2 高频关键词（言语行为动词）词频统计

序号	关键词	频次	序号	关键词	频次	序号	关键词	频次
1	affect	18	13	explain	10	26	focus on	4
2	present	17	14	discuss	9	27	highlight	4
3	use	17	15	have impact on	8	28	assess	3
4	provide	14	16	show	7	29	evaluate	3
5	be	14	17	suggest	7	30	relate (to)	3
6	examine	14	18	include	6	31	correlate	3
7	identify	13	19	analyze	5	32	employ	3
8	investigate	12	20	test	5	33	advance	3
9	develop	11	21	compare	5	34	outline	3
10	find	11	22	have effect on	5	35	mediate	3
11	propose	11	23	associate	4	36	predict	3
12	influence	10	24	through	4	37	help (to)	3
			25	demonstrate	4	38	introduce	3

4.3 亮点的分类

根据已统计的关键词含义对亮点进行分类,再通过时态、语态或形容词对已确定的亮点分类进行确认,具有一定的科学性。本文基于亮点的表现形式和亮点的表达内容两个视角将学术论文中的亮点划分为研究创新型亮点、研究方法型亮点、研究过程型亮点和研究结论型亮点四种类型,这种分类方法具有较好的包容性和继承性。在表 3 中列举每一类亮点下的部分高频关键词及亮点示例:

表 3　亮点分类及示例

亮点类型	特征词(高频)	亮点示例
研究创新型亮点	动词: develop explore, suggest devise, find, propose, present argue, advance, provide 名词:finding perspective	• ***Developing*** the E-health synergy concept to enhance the relationship between IT-enabled resources and hospital performance. • We ***present*** a directory based framework for incentives management of mobile device resources in ad-hoc mobile cloud environment. • ***Advancing*** our theoretical and practical understanding of E-health can be effectively integrated to the realization of E-health strategy. • We ***propose*** a big data architecture to suit the internet of things in data-information and information-knowledge layers. • The ***findings*** suggest that social presence is formed through machine interactivity, person interactivity, and self-disclosure.
研究方法型亮点	动词:use, through, employ, utilize 名词:method approaches, Methodology	• The comparative salience of restaurant attributes is explored by ***employing*** a conjoint analysis. • We ***use*** structuralism and functionalism paradigms to analyze the origins of big data applications. • We ***utilized*** voluntary customer reviews from the smart tourism system. • This paper presents the state of research and main trends in public service management ***through*** a bibliometric analysis.
研究过程型亮点	动词:compare introduce describe outline summarize review, focus on highlight, affect measure, discuss analyze, explain examine, evaluate pay attention to emphasize, assess Investigate 名词: description review, condition factor, drivers analysis, issue technique Application process	• We ***investigate*** the utilization of Facebook by local Korean governments for tourism development. • We ***analyze*** major challenges with big data and also discussed several opportunities. • We ***examine*** the correlation between firm's financial records and vulnerabilities. • A total 110 studies ***reviewed*** to clarify social commerce concept using per-defined review protocol. • We ***highlighted*** the research themes that have been addressed in previous studies. • ***Outlining*** a number of potential research issues in this field of study. • We provide a description of existing communication technologies used in smart cities.
研究结论型亮点	动词: Demonstrate, validate, identify Enhance, increase Improve, indicate Illustrate, define result in, lead to induce, predict will, determine associate (with) relate (to), address relevant(to) 名词:result, trend, explanation	• It ***identifies*** scientific gaps that can promote and guide new studies on improving the existing theory or proposing innovative models. • Both perceived utilitarian and social value of a social shopping website ***lead to*** purchase intention. • Case studies and emerging technologies for big data problems are ***discussed.*** • Financial records ***are significantly associated with*** the number of vulnerabilities. • The ***results*** of the review highlighted the limitation and the gaps in the previous studies in three main aspects • We ***defined*** "IT productivity variance" and focused our effort on it in this paper.

(1)研究创新型亮点。这类亮点描述了研究者针对研究问题的新观点或新发现,与既有成果有显著的不同和实质性进步,是一篇论文中最有价值的内容。此类亮点的句子主要使用:"提出了""发现了""设计了""改进了""给出了""发现"和"观点"等关键词。创新是一篇论文的灵魂[15],因此每一篇具有研究成果的论文都应该存在研究创新型亮点。

(2)研究方法型亮点。这类亮点是对作者在论文中明确提出的解决研究问题所采取的针对性方法的简要介绍。研究方法对于解决特定问题具有一定的新颖性和创新性,是具体实施的方法,因此,这类亮点通常不是泛泛地描述一般科学研究方法(如观察法、实证研究法、调查问卷法、专家访谈法等)和问题解决办法(如计量法、共现法、聚类法等)。此类亮点的句子主要使用:"使用了""利用了""通过""方法"和"途径"等关键词。

(3)研究过程型亮点。这类亮点主要描述论文研究过程中获得的成果,这些成果虽然创新性不及研究创新型亮点中描述的显著创新成果与发现,但也可以推动既有研究理论的改进与发展。由于一篇学术论文中描述研究过程的篇幅占比例最大,因此研究过程型亮点的数量也是四种亮点中最高的,此类亮点的句子主要使用:"比较了""讨论了""分析了""评估了""概述了""检查了""调查了"和"强调了"等关键词。

(4)研究结论型亮点。顾名思义,这类亮点是对有价值的研究结论进行阐述。一篇学术论文的基本逻辑是采用了某种方法,进行了一系列的研究,最终得到了某些研究结论,因此这类亮点往往是对研究方法型

亮点和研究过程型亮点的继承和总结。此类亮点的句子主要使用:“实现了”“阐述了”“定义了”“得出了”“提升了”“加强了”“说明了”和“导致了”等关键词。研究结论的描述不一定具有突出的创新型,得出研究结论只是陈述研究过程的最后一步。

此外,基于语篇语义学研究视角,可以通过时态、语态或形容词来重新确认亮点的分类,如表 4 所示:

表 4 基于时态、语态和形容词的亮点示例

时态(Tense)	• The phase lag index (PLI) ***was used to*** assess local and large-scale connectivity.(研究方法型亮点) • Experts in a variety of taxa ***scored*** five dimensions of intelligence.(研究结论型亮点) • It ***is being increasingly used for*** adults with refractory epilepsy.(研究过程型亮点) • Increasing pulse voltage ***will*** increase particle reduction efficiency.(研究过程型亮点)
语态(Voice)	• miRNA microarray ***was performed*** on cortical dysplasia and compared with normal.(研究方法型亮点) • We ***discuss*** the changes of balance-related variables during static standing.(研究过程型亮点)
形容词(Adjectives/Adverbs)	• There was a ***stronger*** link between human capital of common workers and labour productivity.(研究结论型亮点) • Of somatic comorbidities, stroke showed the ***strongest*** association with epilepsy.(研究结论型亮点) • The algorithm is ***capable*** of optimising stochastic and uncertain problems.(研究结论型亮点) • Fribe optic technology is ***effective*** for measuring spinal curvature over large regions of the spine.(研究过程型亮点) • Restricting participation to the ‘directly affected’ is ***far too*** narrow.(研究过程型亮点) • Clarifies discussion on slurs by introducing ***new*** distinctions and terminology.(研究过程型亮点) • We recommend detailed information for ***future*** designed protocol.(研究结论型亮点) • For some complicated and sensitive cases like nuclear energy, conducting a RSIA is ***necessary***.(研究结论型亮点)

通过以上亮点的语言学特征分析,基本确定了各类型亮点的常用表达方式,为后续亮点的特征项抽取研究打下了基础,例如在抽取研究方法型亮点时,选取文中的“use”“through”“method”等词;在抽取研究创新型亮点时,选取“finding”“perspective”“explore”等词。

5 亮点在论文内的位置分布规律分析

亮点的分布特征是指亮点在全文和文章各部分中是如何分布的。每类学术论文都有一定的逻辑结构,其中不同部分的亮点往往具有不同的动机和功能。因此本文的研究思路是,首先分析学术论文的结构,然后依据上一部分的分析,对每一部分亮点出现的类型与数量进行统计分析。

5.1 论文结构分析

学术论文普遍采用 IMRAD 规范结构。由引言(介绍研究背景和提出研究问题)、材料与方法、结果和讨论四部分构成。对于不同的学科,IMRAD 结构存在许多变体,例如在数据驱动型学科中,“材料与方法”相应的改成“数据与方法”。本研究需要结合论文普遍规范结构对目标期刊 *International Journal of Information Management* 中学术论文的结构进行调查统计,统一本研究中论文采用的结构规范。

5.1.1 全文的章节分布

调查发现,385 篇期刊论文中,四至六节式论文共有 264 篇,占全部样本的近七成。其中,出现最多的是五节式论文的结构,共有 110 篇,占28.6%;其次是四节式,共有 79 篇,占 20.5%;然后是六节式论文,共 75 篇,占 19.5%。其他结构式论文合计占三成左右,而且其中还有相当一部分论文不属于完全研究型论文(见图 3)。

5.1.2 文章各部分的标题内容

研究发现,文章各部分的标题内容呈现出较高的多样性。由于在一篇学术论文中普遍存在多个部分具有相同结构功能的现象,因此需要对各部分标题进行人工判断、甄别、合并和归类,最终形成一个相对统一的论文结构,以便进一步对亮点的分布位置进行解读。根据各部分标题的频数统计情况,绘制了论文各部分名称结构图(见图 4),从左到右依次为四节式论文、五节式论文和六节式论文,按照从图的上端到下端的次序,依次呈现论文各个部分的标题,各部分由白线划分。在每一部分中,矩形的大小表示标题频数的高低,并按照赤橙黄绿青蓝紫的顺序从多到少依次着色。

5.1.3 统一论文结构

由于期刊论文结构和各部分名称的多样性,必须要通过总结归纳,制定一个相对统一的论文结构,可以将大部分论文内容嵌套进去。因此需要秉持“化繁为简”的原则,根据具体情况进行归并。本文通过分析确定了“Introduction(引言)-Research(研究工作)-Method/Methodology(方法)-Results(研究结果)-Conclusion(研究结论)”的五节式论文结构。对于案例分析、综述和述评等其他结构不够规范的文献,按照三段式结构处理:“Introduction-Research-Conclusion”,将中间论述的多个章节的分主题都归为“研究工作”。

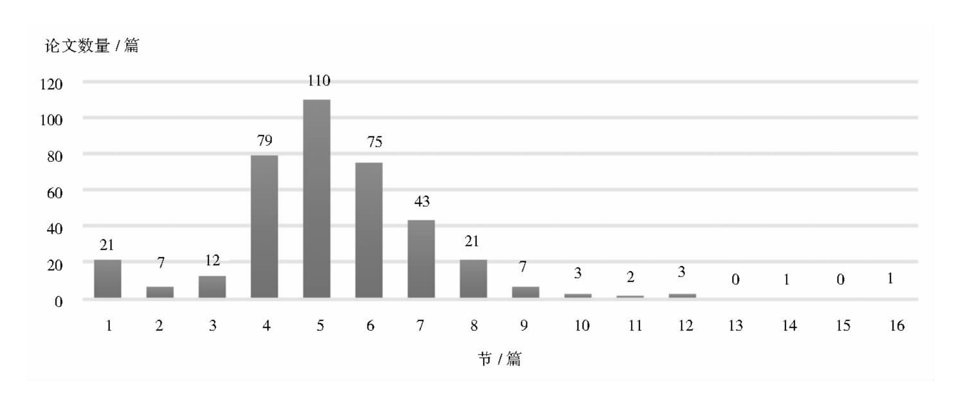
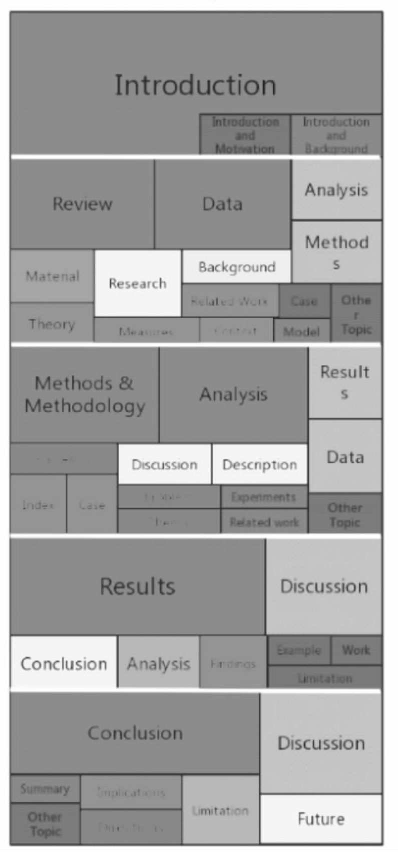


图 3 论文节数分布





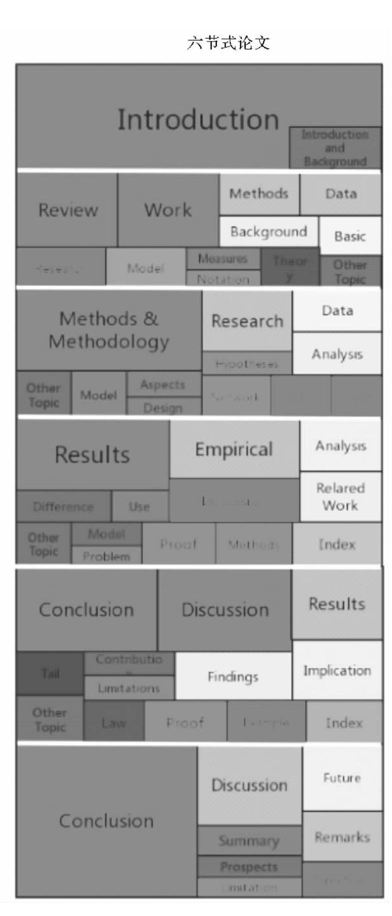


图 4 论文章节结构图及标题名词

5.2 XML 文本数据的解析

本文使用 Python 程序对语料库进行解析，在 Python 解析 XML 的常用方法中，“xml. etree. ElementTree”模块（简称 ET），具有方便友好的 API，且代码可用性好、速度快、消耗内存少。因此本研究中选用该方法进行语料库解析。下面以读取文献 17 为例，结合部分 python 代码对 XML 文件处理过程进行介绍（见表 5）。

解析文件的思路是将 xml 文件的内容看作一个树形结构，它是由一层一层节点分散组成的，例如在本研究中，根节点为“<publication>”标签，第一节子节点分别为“<journal>”“<time>”“<title>”“<author>”“<doi>”“<highlight>”“<abstract>”“<keywords>”和“<section>”等。“<abstract>”和“<section>”的第二节子节点是描述亮点与论文中文字匹配情况的“<h>”标签，所以要得到或操作各个节点的值，就需要依次进行遍历操作。而后，通过获取二级子节点的标签、属性和本文值可以清楚地查看亮点地匹配情况和具体内容，并加以人工统计，为探索亮点的位置分布提供了有力的数据支持。

表 5　XML 文本处理示例

```
>>> import xml.etree.ElementTree as ET #遍历文件
>>> tree = ET.parse(r"C:\Users\…文件路径….xml")
>>> print(tree)
>>> print(type(tree))
>>> root = tree.getroot() #得到根节点
>>> rtag = root.tag #根节点的标签
>>> print(rtag)
publication
>>> print('root_tag:{}'.format(root.tag))
root_tag:publication
>>> print("root_attrib:{}".format(root.attrib))
root_attrib:{'marker':'yu guoxin'}
>>> for i in root:#遍历根节点,得一级子节点
    >>> ctag = i.tag #获取一级子节点的标签
    >>> print(ctag)
    >>> print(type(ctag))
Journal <class 'str'>;time <class 'str'>;title <class 'str'> #输出部分结果
author <class 'str'>;doi <class 'str'>;highlight <class 'str'>
abstract <class 'str'>;keywords <class 'str'>;section <class 'str'>
……
    >>> catt = i.attrib #输出部分结果
    >>> print(catt)
{'name':'Conclusions','category':'conclusion'}
    >>> print(type(catt)) #dict 字典组成的键值对
>>> for j in i:#遍历二级子标签
    >>> jtag = j.tag #获取二级子节点的标签
>>> print(jtag)
h;h #文中两处亮点匹配的部分
    >>> jatt = j.attrib #获取二级子节点的属性
    >>> print(jatt)
{'target':'4','match':'full'} #与第四条亮点完全匹配
    >>> jtext = j.text #获取二级子节点的值,若无则为 None
    >>>print(jtext)
Finally, we outlined several open research challenges, which must be addressed to improve the overall QoL, user perception, and acceptability of m-learning environments. #显示亮点内容
>>> for i in root.iter("Item"):#查询某种所有类型的标签
    >>> print(i.tag,i.attrib,i.text)
>>> print (root[0].text) #下标访问各个标签、文本
>>> print (root[1][1][0].text)
```

5.3　亮点在论文内的位置分布规律

5.3.1　亮点在正文的分布情况

本文所调查的 385 篇论文共有 1 649 条亮点，由于存在一条亮点在论文中匹配多次的情况，于是这些亮点分布在正文四千多处(见图 5)。其中，引言(Introduction)部分出现亮点 602 次；研究工作(Research)部分出现亮点 325 次；研究方法(Method/Methodology)部分出现亮点 873 次；研究结果(Result)部分出现亮点 1 472 次；研究结论(Conclusion)部分出现亮点 810 次。

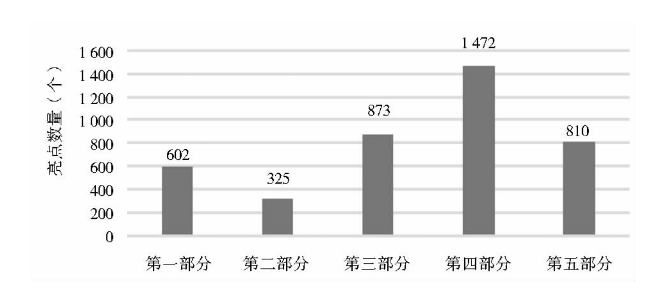


图 5　亮点正文分布数量

在图 6 中可以清晰看出不同类型的亮点在正文中的分布特点并不相同。①引言部分是论文的概述，通常会用简练的语言描述一篇论文的研究问题、解决问题的方法、重要研究成果以及研究结论等，因此引言部分文字通常会匹配全部类型的亮点。②研究工作部分描述研究实施过程，因此该部分主要包含研究过程型亮点。③研究方法部分主要描述具体的研究方法和研究方法实施的过程，因此主要包含研究方法型亮点与研究过程型亮点。④研究结果是一篇论文的核心部分，体现一篇论文的创新性，因此主要包含研究创新型亮点。⑤研究结论部分主要包含研究结论型亮点，同时该部分也会创新性的对既有研究理论进行升华，会对论文中的部分研究过程进行复述，因此该部分也包含研究创新型亮点和研究过程型亮点。

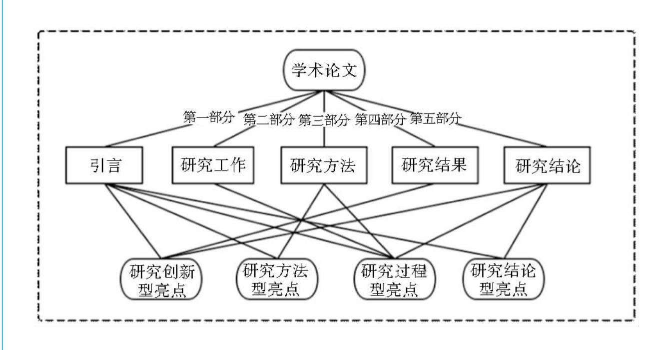


图 6　各类型亮点正文分布

5.3.2　亮点在各部分内的分布情况

除了正文中的分布情况，本文还分析了亮点在各部分的分布规律和特点。比如，亮点倾向于出现在某一部分的开头部分，还是结尾部分，而不同位置的亮点

又意味着什么。

为探索亮点在各部分内部的分布特点,将论文的每一部分内容拆分为前部、中部和后部。调查发现,在各部分的内部,亮点位置的分布是随机的,没有在前部、中部和后部显著聚集的特点。唯独在第一部分“引言”中,由于后部的位置一般用来陈述论文的研究目标和研究路径,所以亮点出现的次数较少,且本研究将研究背景和相关研究综述统一归并到“引言”部分中,因此,在“引言”部分中,亮点大多分布在前部和中部。

6 结论

长期以来,如何快速高效地发现学术论文中有价值的内容片段,从而推动知识创新的速度与质量,一直都是图书情报与学术出版领域的科学难题。目前国内外学者主要是基于文本内容挖掘视角开展了一些研究,例如文献 18、19 与 20 等。爱思唯尔亮点的提出,推动了这项研究的发展。本文对亮点的概念做了进一步的界定,并对其语言学特征和在论文中的位置分布规律进行了分析,研究结果如下:①学术论文研究亮点是一组规范的、语义明确的短句集合。具有新颖性、简明性、易读性、“宣传”性的显著特点。②亮点对于读者、审稿编辑、期刊出版商和作者均有重要意义和价值,亮点有利于更高效地进行论文审稿,助力学术监审;有利于提高学术出版商论文的吸引力,拓展增值效益;有利于读者对论文价值的判断,提高阅读效率;有利于论文作者宣传自己的论文,传播学术观点。③亮点可以分为研究创新型亮点、研究方法型亮点、研究过程型亮点与研究结论型亮点。④亮点分布在论文的各个部分,主要分布在研究结果部分与研究方法部分,在各章节中呈现无序的随机分布。

最后,由于语料库的原因,本文仅选取图书情报领域的一种英文期刊的 385 篇论文进行了语言学特征和位置分布特征分析,其结论存在一定的局限性。后续将选择更多学科与期刊,以及汉语等不同语言的学术论文来丰富语料库,完善对亮点的特征分析的进一步探索。从而为制定亮点自动抽取规则提供更加科学的依据。

参考文献:

[1] Elsevier. Research highlights[EB/OL]. [2018-11-18]. https://www.elsevier.com/authors/journal-authors/highlights.html.
[2] HYLAND K, GUINDA C S. Stance and voice in written academic genres[M]. England: Palgrave Macmillan UK, 2012.
[3] YANG W, WENHSIEN. Evaluative language and interactive discourse in journal article highlights[J]. English for specific purposes, 2016, 42:89-103.
[4] RONZANO F, SAGGION H. Knowledge extraction and modeling from scientific publications[J]. International workshop on semantic, analytics, visualization, 2016(9792):11-25.
[5] DAHL T. Contributing to the academic conversation: a study of new knowledge claims in economics and linguistics[J]. Journal of pragmatics, 2008, 40(7):0-1201.
[6] FISAS B, SAGGION H, RONZANO F. On the discoursive structure of computer graphics research papers[C]//Proceedings of the 9th linguistic annotation workshop. Denver: Association for Computational Linguistics, 2015:42-51.
[7] 温有奎, 吴广印. 碎片化科研创新点动态挖掘研究[J]. 数字图书馆论坛, 2014(7):25-32.
[8] 乐小虬. 领域内中文科技文献中新发现语言描述特征分析[J]. 现代图书情报技术, 2016(5):47-55.
[9] 温有奎,吴广印. 碎片化科研创新点动态挖掘研究[J]. 数字图书馆论坛,2014(7):25-32.
[10] 毛琛瑜,乐小虬. 领域内中文科技文献中新发现语言描述特征分析[J]. 现代图书情报技术,2016, 32(5):47-55.
[11] 李瑛,周立. 科技期刊论文创新点合理呈现的价值及理想模式[J]. 中国科技期刊研究, 2018,29(10):993-999.
[12] SCOTT M, TRIBBLE C. Textual patterns: key words and corpus analysis in language education[M]. Philadelphia: John Benjamins, 2006.
[13] 钟守满,张伟华. 英汉言语行为动词分类及其语义认知解释[J]. 上饶师范学院学报, 2004(5):88-91.
[14] 陈昌来. 现代汉语动词的句法语义属性研究[M]. 上海:学林出版社, 2002.
[15] 李怀祖. 管理学科博士论文撰写探讨[J]. 学位与研究生教育, 2000(3):21-27.

作者贡献说明:

索传军:负责研究课题选题、思路和框架的提出,指导论文写作、修改与完善;

于果鑫:负责研究课题有关数据的处理与分析,论文初稿的写作。

Exploration of the Research “Highlights” in Academic Papers

Suo Chuanjun　Yu Guoxin
School of Information Resource Management, Renmin University of China

Abstract: [Purpose/significance] In the process of publishing a paper, if the core viewpoint of an academic paper can be presented reasonably and effectively, it can not only greatly reduce the time spent by researchers in searching and screening literature, but also help to read and understand. [Method/process] By annotating 385 journal papers in XML format, a research corpus was constructed, and then the linguistic characteristics of highlights were analyzed by keyword analysis method, and the distribution characteristics of highlights were explored by natural language processing algorithm. [Result/conclusion] The highlight of this paper is the collection of a set of normative and clear-cut short sentences, which is the embodiment of new viewpoints, new perspectives, new methods, new ideas, new results and new conclusions in an academic paper compared with other papers. The highlights are novel, concise, readable and “propaganda”. In addition, this paper divides the highlights into research innovation highlights, research methods highlights, research process highlights and research conclusion highlights, and finds the distribution of highlights in the text and chapters.

Keywords: highlights　definition of highlight　value of highlight　linguistic characteristic　distribution characteristic